单词统计特性在情感词自动抽取和商品评论分类中的作用 *

韩彤晖，杨东强，马宏伟

(山东建筑大学 计算机科学与技术学院，济南 250100)

摘 要：单词的统计特征在自然语言处理中具有广泛的应用。针对统计特征对关键词抽取和文本分类精确度的影响，分析了八种常见的统计特征，通过情感词抽取和商品评论分类，研究统计特征在情感分析领域中的作用。情感词提取实验的结果表明，通过结合统计特征与词性，情感词提取的准确率能够达到 76.4%，显著高于基于统计特征或单词词性的情感词提取算法。商品评论分类的测试结果表明，与传统的基于单词的文本情感分类相比，基于统计特征的商品评论分类的准确率提高了 10.8%。利用八种统计特征构造文本向量空间模型，替代基于单词构造文本向量空间模型的方法，能够降低文本向量的维度，具有隐形语义空间(LSA/SVD)的压缩效果，在保证分类结果准确率的前提下有效降低了算法的复杂度，能够替代传统的向量空间模型。

关键词：统计特征；情感词提取；商品评论分类

中图分类号：TP391　　doi: 10.3969/j.issn.1001-3695.2017.09.0913

Using lexical statistical features in extracting sentimental words and classifying product reviews

Han Tonghui, Yang Dongqiang, Ma Hongwei

(*School of Computer Science and Technology Shandong Jianzhu University*, *Jinan* 250100, *China*)

Abstract: The statistical features of words are widely used in Natural Language Processing. This paper summarizes eight types of statistical features, and studies the role of these features in extracting sentimental words and classifying product reviews. Sentiment words extraction result showed that combining these statistical features and PoS tags of words can achieve much higher extraction accuracy than other methods with precision of 76.4%. Product reviews classification results showed that in contrast with sentimental words in constructing the feature space, exclusively using these 8 kinds of statistical features can improve classification precision by 10.8%. Different from the multi-dimensions of lexical elements in the vector space models (VSM) , this paper only employed these 8 types of statistical features in representation of words or documents, which has the ability that can lower the VSM's dimension and can effectively derive the latent semantic space without expensive time and space complexity of SVD calculation.

Key Words: statistical features; extracting sentimental words; classifying product reviews

0 引言

文本情感分析作为自然语言研究领域的热点之一，在舆情分析与控制，商品评论系统中具有重要应用。情感词抽取是文本情感分析的基础，其中，抽取精度和范围是情感词典构造[1]、文本情感分类[2,3]和情感强度计算[4]等应用的基础。以语法规则为基础的情感词抽取算法是一种易于实现的情感词自动抽取算法，其中，Qiu[5]根据单词之间的概率关系，挖掘情感词与主题词的语法联系，同步扩充情感词集合和主题词集合，Liu[6]在语法规则的基础上通过引入语义相似性，改进算法的效率。上述情感词提取方法的范围仅仅局限于形容词，但是，在实际的语言环境中情感词不只局限于形容词。单词统计特征的使用不仅能够打破词性和文本领域依赖性的限制，对不同语种也具有较好的适应性[7]。

本文分析了八种自然语言处理领域常见的单词统计特征，通过情感词抽取和商品评论分类，研究这些统计特征在情感分析领域中的作用。情感词提取结果表明，结合统计特征与单词词性的情感词提取算法的提取精度显著高于其他常用算法。商品评论分类的实验结果表明，以八种统计特征为基础构造的低维向量空间模型能够提高分类器的准确率并能够有效降低分类算法的时间和空间复杂度。

基金项目：国家教育部人文社会科学研究一般项目基金资助项目（15YJA740054）

作者简介：韩彤晖（1991-），男，山东滨州人，硕士研究生，主要研究方向为自然语言处理（kevinsdj@sina.com）；杨东强（1970-），男，副教授，博士，主要研究方向为人工智能；马宏伟（1969-），男，教授，博士，主要研究方向为计算机网络应用．

1　相关研究

单词的统计特征以数值的形式反映单词同文类型之间的关联性，这种关联性能够作为提取关键词的依据。点互信息[8](PMI: pointwise mutual information)算法是一种典型的基于统计特征的情感词提取算法，该算法还能够根据被提取单词的 PMI 值来判断单词的情感极性。Aliaksei[9]将语料库中出现的单词作为特征，利用线性分类算法对测试文本进行情感分类，通过调整参数向量优化分类结果，当分类结果达到最优时，将每个特征对应的参数作为提取情感词的标准。Yu [10]认为情感词对文本情感极性的贡献值远大于非情感词，因此在已知文本情感极性的前提下，计算单词在文本内的权重，并根据权重提取情感词。上述算法的实现过于复杂，且基于 PMI 的情感词提取算法在个别领域的文本可靠性不强。

文本情感分类根据单词在文本中的分布特征训练分类模型，可以将情感词作为特殊的关键词构造文本的向量表达。因此，基于统计特征的关键词提取算法同样适用与抽取情感词，例如，Rajeswari[11]将信息增益作为提取关键词的依据，Uysal[12]使用信息增益、让步比、基尼系数在文本中提取关键词。McAuley[13]在 LDA 模型的基础上挖掘文本内的潜在关键词，Chen[14]通过 LDA 模型在文本中挖掘关键词，并根据关键词的频率分布对其进行分类处理。Mesleh[15]使用卡方测试为单词赋予权重，并根据权重提取关键词，Mitra[16]将单词与文本之间的相关系数作为特征提取的主要依据，Juola[17]利用交叉熵计算文本同单词之间的关联性，根据关联强度挖掘关键词。虽然单词的统计特征能够直观的反映单词同文本类型之间的关联程度，但是基于统计特征提取关键词面临着阈值确定的问题。

本文分析了八种常见的统计特征在情感词抽取和文本情感分类中的作用。实验中使用基于机器学习的方法进行文本情感分类，以检验基于八种统计特征构造的向量空间模型对分类算法的优化能力。

2　特征值计算与数据表达

本文依次研究信息增益(IG: information gain)、优势比(OR: odds ration)、互信息(MI: mutual information)、对数概率比(LPR: logarithmic probability ratio)、交叉熵(CC: cross entropy)、卡方检测(CHI: chi-squire test)、相关系数(CC: correlation coefficient)和差异性分布(DD: differential distribution)在情感词抽取和商品评论分类中的作用。

2.1　特征值计算

使用 C 表示文本情感类型，$C\in\{pos, neg\}$，pos 为积极性情感，neg 为消极性情感，$P(C)$表示文本的情感类型为 C 的概率，$P(\overline{C})$表示文本情感类型为非 C 的概率，其中，$P(\overline{C})=1-P(C)$。使用字母 T 表示表示文本，字母 w 表示单词，$P(w)$表示在 T 中包含 w 的概率，$P(\overline{w})$表示 T 中不包含 w 的概率，$P(\overline{w})=1-P(w)$。为了便于计算单词的统计特征，本文创建四元组 $Q_w=\langle q_p,\overline{q_p},q_n,\overline{q_n}\rangle$，存放单词的分布信息，其中 q_p 表示包含 w 的积极性文本的频率，$\overline{q_p}$ 表示不包含 w 的积极性文本的频率，$q_p+\overline{q_p}$ 为积极性文本的频率，同理 q_n 与 $\overline{q_n}$ 表示 w 在消极性文本中的分布信息，$q_n+\overline{q_n}$ 为消极性文本的频率，表 1 列举了特征值计算过程中使用到的概率近似公式。

表 1　概率近似公式

类型	近似表达			
N	$q_p+\overline{q_p}+q_n+\overline{q_n}$			
$P(w)$	$(q_p+q_n)/N$			
$P(C)$	$C=pos$：	$(q_p+\overline{q_p})/N$	$C=neg$：	$(q_n+\overline{q_n})/N$
$P(w\|C)$	$C=pos$：	$q_p/(q_p+\overline{q_p})$	$C=neg$：	$q_n/(q_n+\overline{q_n})$
$P(w\|\overline{C})$	$C=pos$：	$q_n/(q_n+\overline{q_n})$	$C=neg$：	$q_p/(q_p+\overline{q_p})$
$P(w,C)$	$C=pos$：	q_p/N	$C=neg$：	q_n/N
$P(\overline{w},C)$	$C=pos$：	$\overline{q_p}/N$	$C=neg$：	$\overline{q_n}/N$
$P(w,\overline{C})$	$C=pos$：	q_n/N	$C=neg$：	q_p/N
$P(\overline{w},\overline{C})$	$C=pos$：	$\overline{q_n}/N$	$C=pos$：	$\overline{q_p}/N$
$P(C\|w)$	$C=pos$：	$q_p/(q_p+q_n)$	$C=neg$：	$q_n/(q_p+q_n)$
$P(C\|\overline{w})$	$C=pos$：	$\overline{q_p}/(\overline{q_p}+\overline{q_n})$	$C=pos$：	$\overline{q_n}/(\overline{q_p}+\overline{q_n})$

1) 信息增益(IG)

单词的信息增益表示单词携带的用于区分文本情感类型的信息量，w 的信息增益越大，则表明其区分文本情感极性的能力越强。IG 的计算公式如下：

$$IG(w)=\left\{-\sum_{C\in\{pos,neg\}}P(C)\times\log P(C)\right\}-\left\{\sum_{t\in\{w,\overline{w}\}}P(t)\times\left[-\sum_{C\in\{pos,neg\}}P(C|t)\times\log P(C|t)\right]\right\}\tag{1}$$

2) 改进的让步比(OR)

让步比反映单词影响文本情感极性的能力，让步比的绝对值越高，表明单词影响文本情感极性的能力越强。OR 的计算公式如下：

$$OR(w,C)=P(w)\times\log\frac{P(w|C)\times\left[1-P(w|\overline{C})\right]}{\left[1-P(w|C)\right]\times P(w|\overline{C})}\tag{2}$$

3) 互信息(MI)

互信息指单词携带的能够反映文本情感类型的信息量，w 的互信息越高，表明其携带的信息量越大。MI 的计算公式如下：

$$MI(w,C)=\log\frac{P(w,C)}{P(w)P(C)}=\log\frac{P(w|C)}{P(w)}\tag{3}$$

最终，w 的 MI 值为 $MI(w)=\max_{C\in\{pos,neg\}}\{MI(w,C)\}$。

4) 改进的对数概率比(LPR)

对数概率比类将单词在积极性和消极性文本中出现概率的对数比值作为衡量单词携带信息量的标准，对数概率比的绝对值越大，w 区分文本情感极性的能力越强。LPR 的计算公式如下：

$$LPR(w)=P(w)\times\log\frac{P(w|C)}{P(w|\overline{C})}\tag{4}$$

5) 交叉熵(CE)

交叉熵用于描述单词在积极性和消极性文本之间的分布差

异，若 w 具有较高的交叉熵，则表明其在两种不同极性文本中的分布差异越明显，成为情感词的概率越高。CE 的计算公式如下：

$$CE(w)=P(w)\times\sum_{C\in\{pos,neg\}}P(C|w)\times\log\frac{P(C|w)}{P(w)} \tag{5}$$

6) 改进的卡方检测(CHI)

卡方检测用于测试单词与文本情感类型之间的关联程度，在计算过程中，假定单词和文本类型之间服从自由度为 1 的卡方分布。卡方值越高，表明 w 与文本情感类型的关联度越高，其成为情感词的概率也越大。CHI 的计算公式如下：

$$\chi^2(w,C)=P(w)\times\frac{(A\times D-E\times B)^2}{(A+E)\times(B+D)\times(A+B)\times(E+D)} \tag{6}$$

其中：A 表示情感类型为 C，并且包含 w 的文本的数量；B 表示情感类型为非 C，并且包含 w 的文本的数量；E 表示情感类型为 C，并且不包含 w 的文本的数量；D 表示情感类型为非 C，并且不包含 w 的文本的数量。最终，w 的 CHI 值为：$\chi^2(w)=\max_{C\in\{pos,neg\}}\{\chi^2(w,C)\}$。

7) 相关系数(CC)

相关系数表示单词和文本情感极性之间的相关程度，相关系数越大，表明单词区分文本情感的能力越强。CC 的计算公式如下：

$$CC(w,C)=\frac{\left[P(w,C)P(\bar{w},\bar{C})-P(w,\bar{C})P(\bar{w},C)\right]}{\sqrt{P(w)P(\bar{w})P(C)P(\bar{C})}} \tag{7}$$

8) 差异分布(DD)

差异性分布将单词在积极性和消极性文本之间的分布差异作为衡量单词情感极性的标准，如果 w 在 C 类文本中的频率明显高于(或低于)其在非 C 类文本中的频率，则表明 w 成为情感词的可能性越大，且 w 的情感极性与文本的情感极性相同(或相反)。DD 的计算公式如下：

$$DD(w)=P(w)\times\frac{|P(pos|w)-P(neg|w)|}{\max_{C\in\{pos,neg\}}\{P(C|w)\}} \tag{8}$$

分母取 $P(pos|w)$与 $P(neg|w)$之间的最大值， DD(w)的取值范围为[-1,1]，乘以概率 $P(w)$的目的是为了降低噪声对 DD 值的影响。

由于标准的让步比、对数概率比和卡方检测算法倾向于给低频率单词赋予较高的权重，使得大量低频非情感词具有较高的权重，从而影响算法的可靠性。为了提高算法的可靠性，本文在标准算法的基础上乘以单词概率 $P(w)$。

以让步比为例，表 2 列举了一组单词的 OR 值以及在特征值列表内的排列顺序。由表 2 可知，让步比算法改进前后，单词在特征值列表中的排列顺序变化较为明显。基于让步比的情感词提取实验表明，与基于标准让步比的情感词提取算法相比，基于改进让步比的情感词提取算法的准确率提高了 17.8%。

上述特征值的计算过程基本相似，本文以计算单词'good'的信息增益为例，介绍特征值的具体计算方法。根据'good'的分布信息，构造四元组 Q_{good}，统计结果显示，包含'good'的积极性和消极性文本频率分别为 9974、5978，不包含'good'的积极性和消极性文本频率分别为 23400、27464，即 q_p =9974、$\overline{q_p}$ =23400、q_n =5978、$\overline{q_n}$ =27464，因此 Q_{good}=<9974,23400,5978,27464>。将 Q_{good} 带入表 1，得到计算'good'信息增益所需的相关概率，计算结果如表 3 所示。

表 2　使用标准让步比计算方法与改进的让步比计算方法计算的部分单词的 OR 值和这些单词在特征值列表内的序号

单词	标准的 OR 算法		改进的 OR 算法	
	OR 值	排序	OR 值	排序
wtf	2.3263	39	0.0016	932
yellow	2.2399	48	0.0017	841
penny	2.1527	55	0.0055	256
great	1.5835	247	0.3461	1
love	2.0303	77	0.3156	2
good	0.6270	1121	0.1604	5

表 3　概率计算结果

类型	取值	类型	取值
$P(good)$	0.2387	$P(pos\|good)$	0.6253
$P(\overline{good})$	0.7613	$P(neg\|good)$	0.3747
$P(pos)$	0.4995	$P(pos\|\overline{good})$	0.4601
$P(neg)$	0.5005	$P(neg\|\overline{good})$	0.5399

将上述概率代入式(1)，得到 IG(good)=1.0003。

2.2 情感词自动提取

情感词的提取过程的实质是对连续型特征做离散化处理，跟据统计特征将单词分配到情感词或非情感词集合中。为了快速合理的划分单词集合，采用 SDR(standard deviation reduction)[18]算法划分单词集合，确定统计特征对应的阈值。SDR 算法采用动态方式将单词分配到相应的集合中，分配操作结束后，计算该次分配的误差缩减量，当误差缩减量达到最大时，表明分配结果达到最优。算法公式如下：

$$value=sd(L)-\frac{|L_s|}{|L|}\times sd(L_s)-\frac{|L_n|}{|L|}\times sd(L_n) \tag{9}$$

其中：L 表示由候选情感词的特征值组成的列表，L 中的元素按照特征值递减的顺序排列，L_s 表示情感词特征值列表，L_n 表示非情感词特征值列表，$L=L_s+L_n$。|•|表示集合或列表中元素的数量，sd(•)为标准差函数。当 $value$ 达到最大值时，对情感词和非情感词的划分达到最优，此时 L_n 内的最大特征值即该统计特征对应的阈值。算法 1 描述了 SDR 的执行过程。

为了演示 SDR 的执行过程，以信息增益为例，创建包含 10 个信息增益的样本列表 L，通过 SDR 确定样本的阈值。在 L 中按信息增益递减的顺序排列，如表 4 所示。首先将样本集合 S 划分为两个列表，即情感词特征值列表 L_s(={$IG_{disappoint}$})和非情感词特征值列表 L_n(={IG_{happy}~IG_{well} })，在 L_s 和 L_n 中，按元素值递减的顺序排列。在算法执行过程中，若 $value>V$ 时，使用 $value$ 更新变量 V，并且，通过 L_n 中最大的特征值更新变量 $threshold$。当算法执行结束时，$threshold$ 的取值就是在 L 中抽取情感词的阈值。

算法 1

```
输入：特征列表 L，情感词特征值列表 L_s 和非情感词特征
      值列表 L_n，其中，L={f_w1~f_w|L|}、L_s={f_w1}、L_n={f_w2~f_w|S|}
过程：
    创建变量：V=-1，threshold =∀；
    for  |L_n| > 1  do
        计算 L、L_s 和 L_n 的标准差 sd(L)、sd(L_s)和 sd(L_n)；
        将 sd(L)、sd(L_s)和 sd(L_n)带入公式(12)，计算 value；
        if  value > V  then
            V=value，  threshold= max_{wi∈s_n} { f_wi } ；
        end  if
        f_max= max_{wi∈s_n} { f_wi } ；
        将 f_max 存入列表 L_s，并在列表 L_n 中删除 f_max；
    end  for
输出：threshold
```

计算 L、L_s 和 L_n 的标准差和列表长度，得到，样本整体的标准差 $sd(L)\approx0.2254$，长度$|L|=10$；情感词列表的标准差 $sd(L_s)\approx0.0$，列表长度$|L_s|$为=1；非情感词列表的标准差 $sd(L_n)\approx1.861$，列表长度$|L_n|$为 9。

计算得到 $value\approx0.0579$，将 $value$ 赋值给 V，将 L_n 中的最大值 IG_{happy} 添加到 L_s，即 L_s={ $IG_{disappoint}$，IG_{happy} }，并且 $threshold$=IG_{happy}，最后，在 L_n 中删除 IG_{happy}，得到 L_n={ $IG_{perfect}$~IG_{well} }。程序的最终执行结果显示，该信息增益样本的最佳阈值为 0.8053。

表 4　单词样本

单词	IG	单词	IG
disappoint	1.3370	best	0.7816
happy	1.2387	fast	0.7441
perfect	1.1125	awesome	0.7276
good	1.0003	stop	0.7266
amaze	0.8053	well	0.6931

本文对被提取的单词做如下定义：若基于单一统计特征 α(∈{IG,OR,MI,LPR,CE,CHI,CC,DD})提取情感词，α 对应阈值为 θ，w 关于 α 的特征值为 f_w。若 $f_w>\theta$，则认为 w 是情感词，称 w 满足统计特征 α。

除了测试使用单一特征提取情感词的效果，本文还测试了基于多统计特征的情感词提取方法。实验根据研究的统计特征的数量设置了 8 种提取标准，依次为 C_1~C_8，其中，C_i($i\in$[1,8])要求被提取的单词至少满足 i 种统计特征。

2.3　情感分析中的数据表示

基于 2.1 中介绍的统计特征创建单词的特征向量，实现单词的向量表示，向量的格式如下：

$$\boldsymbol{V_w}=\{f_1,f_2,f_3,f_4,f_5,f_6,f_7,f_8\}$$

其中：向量元素 f_1~f_8 依次对应八种统计特征。

利用语义组合中的向量加函数构造向量空间模型，通过特征向量表示文本，向量空间模型的构造方式下：

$$\boldsymbol{V_T}=\sum_{u=1}^{t} sig(w_i)\times\boldsymbol{V_w^u} \tag{10}$$

其中：w_i 表示在情感词中编号为 i 的单词号，$sig(w_i)$为符号函数，当 w_i 在 T 中出现时 $sig(w_i)$=1，否则 $sig(w_i)$=0，$\boldsymbol{V_w^i}$ 表示 w_i 对应的单词向量，最终，可以通过向量 $\boldsymbol{V_T}$ 表示文本，在 4.3.2 中详细介绍了文本向量的构造过程。

3　情感词提取与商品评论分类

虽然中文购物网站提供了大量商品评论，但是，现有的中文分词工具一定局限性[19]，并且这些评论中广告信息和虚假评论比重较大。因此，采用中文商品评论难以有效验证算法的实际效果。英文评论能够降低分词错误对算法的影响，并且，亚马逊购物网站提供的英文商品评论信息相对也更加真实。因此，实验采用亚马逊英文网站提供的商品评论，其中积极性评论 33374 条，消极性评论 33442 条。

3.1　文本预处理

在网站内采集的商品评论包含大量停止词和单词缩写，因此，需要对这些数据进行预处理，操作步骤如下：

a) 文本规范化(normalization)。将大写字母转换为小写字母，过虑特殊符号(如：#、@)和停用词(如：this、that)，将单词缩写替换为正规格式(如：that's→that is)，将否定性副词统一替换为 not(如：hardly→not)。

b)词干处理(stemming)。若单词以名词复数、形容词比较级、动词过去式等形式出现，则将该单词还原(如 issues→issue、better→good)。

c) 词组抽取。多个连续的单词之间存在语法联系，使得这些连续的中性词具有表达情感的能力，如 meet my expectation、not buy again 等。本文根据单词间的语法联系抽取情感词组。

d)候选词列表构造：构造候选情感词列表 L，将文本中出现的单词和短语作为候选情感词存储在列表 L 中，L 内不包含重复出现的单词和短语；

e)频率限制：过虑频率低于 β 的单词，本实验将 β 设置为 35；

表 5 展示了预处理前后，评论集合中单词总量以及形容词、动词、名词、副词的数量变化。

表 5　预处理前后语料库中单词的数量变化

	预处理之前		预处理之后	
	word token	word type	word token	word type
形容词	464686	29443	206331	512
动词	1009911	29513	85451	228
名词	1086026	30281	630252	1596
副词	431654	27807	105066	191
积极类文本的词汇	2855588	25121	633362	2758
消极类文本的词汇	2378824	23907	546016	2758
语料库的词汇	5234412	34328	1179378	2761

表 6　部分单词及其特征分布(加粗部分表示满足阈值要求)

单词	词性	IG	OR	MI	LPR	CE	CHI	CC	DD
		θ_{f1}=0.1017	θ_{f2}=0.0065	θ_{f3}=0.0022	θ_{f4}=0.0107	θ_{f5}=2.7921	θ_{f6}=0.0038	θ_{f7}=7.4363	θ_{f8}=0.0063
luck	n	**0.1033**	**0.0077**	**0.0050**	0.0076	2.6904	0.0019	**11.3849**	0.0043
awful	adj	**0.1214**	**0.0084**	**0.0056**	0.0083	2.6252	0.0021	**12.2506**	0.0043
cheap	adj	0.0164	**0.0069**	**0.0032**	0.0067	**9.3314**	0.0019	4.6780	0.0060
refuse	v	0.0814	**0.0066**	**0.0042**	0.0066	2.6427	0.0015	**10.1777**	0.0039
nice	adj	**0.3661**	**0.0394**	**0.0152**	**0.0376**	**11.4454**	**0.0558**	**21.8535**	**0.0258**
great	adj	**4.6086**	**0.3461**	**0.0974**	**0.2769**	**21.7177**	**3.2030**	**76.6667**	**0.1569**
love	v	**5.0430**	**0.3156**	**0.0833**	**0.2748**	**22.4717**	**2.3952**	**78.6123**	**0.1289**

3.2　提取情感词

通过四元组计算单词在文本中出现的概率、文本的情感类型为积极或消极的概率、单词同文本情感类型之间的联合概率和条件概率。创建特征值列表 L_{f1}~L_{f8} 依次存储单词和单词的 8 种类型的特征值，并且在特征值列表中单词按照特征值递减的

顺序排列。调用 SDR 算法，计算每一类统计特征的阈值，并创建变量 θ_{f1}~θ_{f8}，依次存储八种统计特征对应的阈值。该实验将基于单词词性的情感词提取算法作为实验基线，其中该算法的准确率为 54.5%，召回率为 27.7%。

3.2.1 基于单一统计特征提取情感词

根据选取的统计特征 α 创建情感词典 D_α 用于存放基于 α 提取的情感词，查找该统计特征对应的特征值列表和阈值，遍历特征值列表，比较列表内单词的特征值与阈值之间的数值关系，若单词的特征值大于阈值，将该单词存入词典，当遍历结束后，词典内的单词即为被提取的情感词。

以基于信息增益的情感词提取算法为例，创建词典 D_{IG}，并将词典初始化为空，信息增益对应的特征值列表为 L_{f1}，阈值为 θ_{f1}=0.1017。遍历 L_{f1}，并比较列表内单词的信息增益与 θ_{f1} 之间的数值关系，若单词的信息增益大于 θ_{f1}，即该单词满足信息增益，将该单词存入词典 D_{IG}。由表 6 可知，单词’luck’的信息增益 IG_{luck}=0.1033，单词’refuse’的信息增益 IG_{refuse}=0.0814。由于 $IG_{luck}>\theta_{f1}$，将’luck’存入 D_{IG}，由于 $IG_{refuse}<\theta_{f1}$，因此’refuse’被过滤。

3.2.2 基于多统计特征提取情感词

基于多统计特征提取情感词，要求被提取的单词满足至少。根据 3.2 的提取标准 *C_1~C_8* 结合提取情感词，并利用 8 种统计特征为被提取单词创建向量表达。算法 2 展示了基于多统计特征的情感词提取算法的执行过程。

算法 2

输入：特征值列表 L_{f1}~ L_{f8}；候选情感词列表 L；阈值变量 θ_{f1}~θ_{f8}

过程：

1. 根据提取标准 *C_1~C_8* 创建词典 D1~D8；
2. **for** $w \in L$ **do**
3. 　$I = 0$;
4. 　查找 w 在 L_{f1}~ L_{f8} 内对应的 8 种特征值 $f_1^w \sim f_8^w$；
5. 　**for** $i = 1,2,\cdots,8$ **do**
6. 　　**if** $f_i^w > \theta_{fi}$ **then**
7. 　　　I += 1
8. 　　**end if**
9. 　**end for**
10. 　**if** $I = n\ (n \in [1,8])$ **then**
11. 　　利用 $f_1^w \sim f_8^w$ 构造 w 的特征值向量 V_w;
12. 　　将 w 和 V_w 存入词典 D1~Dn；
13. 　**end if**
14. **end for**

输出：词典 D1~D8

以提取单词’refuse’的过程为例，介绍该算法。程序从候选情感词列表中读取’refuse’，并将变量 I 初始化为 0。遍历特征值列 L_{f1}~ L_{f8} 表查找’refuse’的 8 种特征值 $f_1^{refuse} \sim f_8^{refuse}$。分别将 $f_1^{refuse} \sim f_8^{refuse}$ 与对应的阈值 θ_{f1}~θ_{f8} 进行数值比较，在比较过程中，若 $f_i^{refuse} > \theta_{fi}$（$i \in [1,8]$），则变量 I=I+1。由表 6 可知，’refuse’同时满足 OR、MI 和 CC，当数值比较结束后，得到 I=3。判断 I>0 是否成立，由于条件成立，因此利用’refuse’的 8 种特征值为其构造单词向量，格式如下：

$$V_{refuse} = (0.0164, 0.0069, 0.0032, 0.0067, 9.3314, 0.0019, 4.6780, 0.0060)$$

将’refuse’和向量 V_{refuse} 存入词典 D1~D3，并从候选词列表中删除’refuse’。之后，程序判断候选情感词列表是否为空，若列表非空，从候选情感词列表中抽取单词，并逐步判断该单词是否为情感词，否则，结束程序。

3.2.3 结合统计特征与单词词性提取情感词

结合单一统计与单词词性的提取算法要求单词的特征值大于对应阈值且单词为形容词，若满足条件则将单词存入情感词典。如表 6 所示，’luck’的信息增益均大于 θ_{f1}，因此该单词满足信息增益，但是，由于’luck’的词性为名词，不满足词性要求，’luck’无法被提取。根据表 6 可知，单词’awful’的词性为形容词，并且 $IG_{awful}>\theta_{f1}$，因此’awful’能够被结提取。

结合多统计特征与单词词性的提取算法的执行过程与算法 2 相似，唯一的区别是第 10 行，不仅要判断 I 的取值，还需要判断候选情感词必须为形容词，如果单词为形容词则继续后面的操作，否则过滤该单词。由表 6 可知，’cheap’和’refuse’都满足三种统计特征，因此这两个单词都能够被算法 2 提取。但是，结合多统计特征与单词词性的提取算法要求被提取的单词必须为形容词，而’refuse’是动词词性，因此’refuse’被过滤，因为’cheap’形容词词性，所以’cheap’能够被程序提取。

3.2.4 提取结果

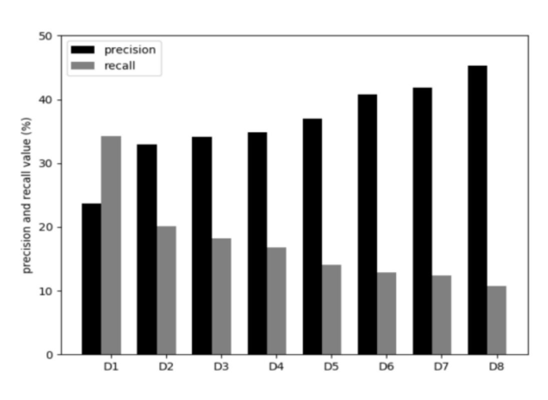
词典 HowNet 为文本情感分类提供了丰富的资源，其中包含 9142 个英文评价词语，本文根据 HowNet 构造标准词典用于检测上述提取算法的效率，标准词典中包含的单词必须在 HowNet 和语料库中同时出现。

表 7 的统计结果表明结合统计特征与单词词性的提取算法具有更高的准确率，相较于基于单一统计特征的提取算法，准确率平均提高 36.8%，而与基于单词词性的情感词提取算法相比，准确率最大提高 21.7%。

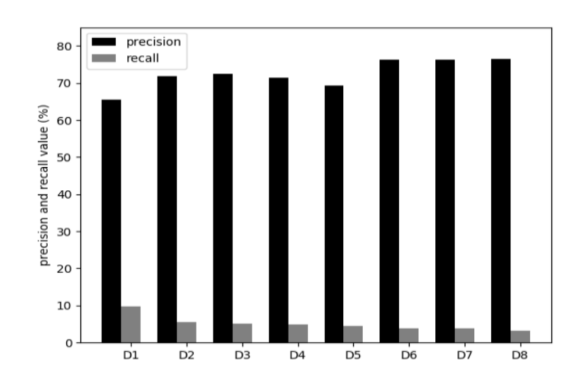
表 7　基于单一统计特征的提取结果

	基于单一统计特征		结合单一统计特征与词性	
	准确率	召回率	准确率	召回率
IG	39.6%	13.3%	70.8%	4.1%
OR	34.4%	17.5%	71.4%	4.9%
MI	33.5%	19.4%	72.4%	5.1%
LPR	40.3%	12.6%	69.6%	3.9%
CE	25.1%	26.7%	67.4%	7.0%
CHI	41.1%	12.9%	76.2%	3.9%
CC	28.9%	23.8%	70.5%	7.5%
DD	37.4%	12.6%	76.2%	3.9%

图 1(a)展示了基于多统计特征提取情感词的结果，横坐标 D1~D8 表示根据 *C_1~C_8* 创建的情感词典，用于存储在对应标准下提取的单词，至少满足一种特征时，提取结果具有最低的准确率，而当八种特征全满足时，提取结果具有最高的准确率。图 1(b)展示了结合多统计特征与单词词性提取情感词，与基于多统计特征的提取算法相比，至少满足一种统计特征时，提取结果的准确率提高了 41.9%，8 种统计特征全部满足时提取结果的准确率提高了 31.1%。



(a) 基于多统计特征的提取结果



(b)结合多统计特征与单词词性的提取结果

图 1 基于多统计特征的情感词提取结果

表 8 单一统计特征的商品评论分类结果(P 表示准确率,R 表示召回率)

	基于单一统计特征(a)										结合单一统计特征与语单词词性(b)									
	朴素贝叶斯		支持向量机		决策树		神经网络		随机森林		朴素贝叶斯		支持向量机		决策树		神经网络		随机森林	
	P (%)	R (%)	P (%)	R (%)	P (%)	R (%)	P (%)	R (%)	P (%)	R (%)	P (%)	R (%)	P (%)	R (%)	P (%)	R (%)	P (%)	R (%)	P (%)	R (%)
IG	85.6	85.6	83.8	83.8	**80.3**	**80.3**	81.5	81.5	81.5	81.5	71.2	70.5	70.6	69.5	70.7	69.6	70.3	69.3	69.9	69.8
OR	86.4	86.4	**85.5**	**85.5**	79.5	79.5	81.5	80.2	83.3	83.3	**71.4**	**70.7**	70.5	69.5	**71.1**	**70.2**	70.2	69.4	70.6	69.8
MI	86.3	86.2	85.4	85.4	79.7	79.7	77.9	77.5	**84.3**	**84.3**	71.4	70.7	70.5	69.5	70.7	70.2	**70.7**	**69.7**	70.7	69.9
LPR	85.4	85.4	84.5	84.5	80.1	80.1	82.4	82.4	82.4	82.4	71.2	70.5	**70.9**	**69.9**	71.0	70.1	70.6	69.8	**71.0**	**70.1**
CE	85.4	85.4	83.6	83.6	79.1	79.1	55.7	55.1	83.8	83.8	70.4	69.9	70.4	69.5	69.1	68.3	66.9	66.3	69.5	68.8
CHI	85.5	85.5	84.6	84.6	79.5	79.5	**83.4**	**83.1**	82.5	82.5	70.5	69.9	70.7	69.8	70.7	69.7	70.3	69.5	70.7	69.9
CC	**87.1**	**87.1**	85.4	85.4	79.6	79.5	75.8	75.8	83.9	83.9	71.4	70.8	70.5	69.5	71.1	70.2	69.5	68.8	70.6	69.8
DD	85.0	85.0	84.6	84.6	78.6	78.6	82.1	80.5	82.8	82.8	70.5	69.9	70.7	69.8	70.7	69.7	70.3	69.5	70.7	69.9

表 9 多统计特征的商品评论分类结果(P 表示准确率，R 表示召回率)

	基于多统计特征(a)										结合多统计特征与语单词词性(a)									
	朴素贝叶斯		支持向量机		决策树		神经网络		随机森林		朴素贝叶斯		支持向量机		决策树		神经网络		随机森林	
	P (%)	R (%)	P (%)	R (%)	P (%)	R (%)	P (%)	R (%)	P (%)	R (%)	P (%)	R (%)	P (%)	R (%)	P (%)	R (%)	P (%)	R (%)	P (%)	R (%)
C_1	**81.4**	**80.2**	**84.1**	**83.9**	**86.7**	**86.7**	**85.9**	**85.8**	**86.3**	**86.3**	70.7	68.5	70.9	69.7	70.6	70.3	70.9	70.0	69.1	68.6
C_2	81.3	80.1	83.5	83.5	84.2	84.2	84.7	84.6	84.7	84.7	70.6	68.4	70.7	69.5	70.7	70.4	70.7	69.4	70.2	69.5
C_3	80.8	79.6	83.5	83.3	84.3	84.2	84.7	84.5	85.0	85.0	70.5	68.3	71.1	69.8	**71.2**	**70.8**	70.5	69.3	70.5	69.9
C_4	81.0	79.8	83.9	83.7	84.5	84.5	85.1	85.0	85.0	85.0	70.6	68.4	71.0	69.7	71.2	70.8	70.6	69.4	70.4	69.7
C_5	80.6	79.3	84.1	84.0	83.9	83.9	85.2	85.1	84.8	84.8	**70.7**	**68.5**	71.0	69.8	70.7	70.3	70.8	69.5	70.5	69.7
C_6	79.9	78.5	83.5	83.4	82.2	82.2	84.4	84.3	84.2	84.2	70.4	68.1	**71.2**	**70.0**	71.2	70.7	70.4	69.0	70.7	70.0
C_7	79.8	78.4	83.4	83.5	82.5	82.4	84.7	84.7	83.5	83.5	70.4	68.1	71.2	70.0	71.2	70.7	70.4	69.0	70.7	70.0
C_8	79.8	78.4	84.1	84.0	82.8	82.8	85.0	84.9	83.4	83.4	70.2	67.8	71.1	70.2	70.7	70.0	**71.1**	**69.9**	**70.9**	**70.1**

3.3 商品评论分类测试

将使用单一特征值构造的情感词典和结合多种特征构造的情感词典用于文本分类测试。实验使用朴素贝叶斯(naïve Bayes)[20,21]、支持向量机(support vector machine,SVM)[20]、决策树(decision tree)[22]、BP 神经网络(BP neural network)[23]和随机森林(random forest)[24]五种算法对测试文本进行情感分类。测试文本同样为亚马逊的商品评论，其中积极评论 997 条，消极评论 999 条，并以列表的形式存放测试文本。使用数据处理工具 Weka 提供的分类器，并使用 10-fold cross-validation 进行商品评论分类测试，每种分类算法的参数均为 Weka 提供的缺省参数。将基于单词词性的商品评论分类结果作为实验的基线，上述五种分类器的测试精度依次为 77.1%、74.5%、69.5%、63.0%和 76.3%。

3.3.1 基于单一统计特征的商品评论分类

将基于单一统计特征构造的情感词典用于商品评论分类测试，并以单词为基础构造向量空间模型。Pang[25]证明，在文本向量中使用 0、1 表示情感词具有更好的效果，因此该实验中文本向量的格式如下：

$$\boldsymbol{SVM}_T=\{sig(w_1),sig(w_2),\cdots,sig(w_t)\}$$

其中:t 表示情感词典中单词的数量，w_i 表示词典中编号为 i 的单词，若 w_i 在 T 中出现，则 $sig(w_i)$=1，否则 $sig(w_i)$=0。

3.3.2 基于多统计特征的商品评论分类

基于词典 D1~D8 对商品评论进行分类测试，该实基于用 8 种统计特征构造向量空间模型，代替传统的文本表方法算法 3 描述了以统计特征为基础构造文本向量的过程。

算法 3

```
输入：文本列表 L_T；词典 Di(i∈[1,8])；
过程：
    创建向量空间列表并初始化为空；
    for  T ∈ L_T  do
      for  (w,V_w) ∈ D1  do
        if  w ∈ T  then
          V_T += V_w
        end  if
      end  for
      将向量 V_T 存入空间向量列表；
    end  for
输出：空间向量列表
```

以对商品评论 T 构造本向量为例，描述文本数据表达的具体细节，构造过程采用的情感词典为 D1，商品评论 T 如下："Very nice, sleek and works great. My grandkids talked me into it, it's what they use in school. So far I love it."其中，'nice'、'great'和'love'以单词向量的形式存储在 D1 中，其向量表示如下：

$$V_{nice}=(0.3661,0.8400,0.0152,0.8025,11.4454,0.0558,21.8535,0.0258)$$
$$V_{great}=(4.6086,1.5835,0.0974,1.2672,21.7177,3.2030,76.6667,0.1569)$$
$$V_{love}=(5.0430,2.0303,0.0834,1.7679,22.4717,2.3952,78.6123,0.1289)$$

遍历词典 D1，依次读取单词和对应的单词向量，检测 $w\in T$ 是否成立，若条件成立，则记录该单词和单词向量。操作结果为：$T\cap D1$={'nice','great','love'}。根据 3.4 可知，$sig(nice)$=1、$sig(great)$=1、$sig(love)$=1、$sig(w)$=0，$w\in D1$ 且 $w\notin T$。得到 $V_T=V_{nice}+V_{great}+V_{love}$，最终，$V_T$ 的向量表达如下：

$$V_T=(10.0177,4.4538,0.196,3.8376,55.6348,5.654,177.1325,0.3116)$$

3.3.3 结合统计特征与单词词性的商品评论分类

结合单一统计特征与单词词性的商品评论分类，以单词为基础创建向量空间模型，文本向量的构造方法与 3.3.1 相同，以 0、1 向量的形式表示文本向量。

结合多统计特征与单词词性的商品评论分类，以 8 种统计特征构造文本的向量表达，尽管该部分文本向量的构造方法与 3.3.2 相同，但是由于情感词提取算法存在差异，因此构造的文本向量也存在差异。仍以 3.3.2 中的商品评论为例，基于多统计特征的商品评论分类通过'nice''great'和'love'，由于结合多统计特征与单词词性的商品评论分类只采用形容词，由表 6 可知，该算法只能利用'nice''great'的单词向量构造评论的向量表达。最终，上述客户评论的向量表达为 $V_T=V_{nice}+V_{great}$，向量形式如下：

$$V_T=(9.6516,3.6138,0.1808,3.0351,44.1894,5.5982,155.279,0.2858)$$

3.3.4 分类结果

表 8(a)展示了基于单一统计特征的商品评论分类结果，在朴素贝叶斯算法中，基于相关系数的商品评论分类算法具最有最高的精度，其准确率为 87.1%。在 BP-神经网络中，基于交叉熵的商品评论分类算法的准确率只有 55.7%。表 8(b)展示了结合单一统计特征与词性的商品评论分类结果，测试结果的准确率集中分布在区间 66%至 72%内。

表 9(a)描述了基于多统计特征的文本情感分类结果，结果表明，当单词至少满足一种统计特时，所有分类算法均达到最优的分类效果。结合多统计特征与单词词性的文本分类的准确率集中分布在区间 69%至 71%内，如表 9(b)所示。

4 实验结果分析

由于人们倾向于用形容词表达个人情感，因此，与基于统计特征的情感词提取算法相比，基于单词词性的情感词提取算法具有更高的精度。在实际的语言环境中，除了形容词，部分动词、副词和名词也具有表达情感的能力，例如，'love'、'kindly'、'issue'等，上述原因使得基于单词词性的商品评论分类算法的准确率低于基于单词统计特征的商品评论分类算法。结合统计特征与词性的情感词提取算法要求被提取的情感词既要满足统计特征并且词性为形容词，表 7 和图 1 展示的提取结果表明，该提取算法的准确率高于只基于单词统计特征或单词词性的提取算法。通过表 8 可知，基于单一统计特征的商品评论分类测试的最高准确率为 87.1%，而结合单一统计特征与单词词性的商品评论分类测试的最高准确率仅为 71.4%。表 9 表明，基于多统计特征的商品评论分类测试的最佳结果为 86.7%，结合多统计特征与单词词性的商品评论分类的最高准确率只有 71.2%。造成上述现象的主要原因是，结合统计特征与词性虽然能够提高提取单词的准确率，但是由于增加了提取算法的限制条件，使得满足要求的单词随之减少，导致文本情感分类测试中情感词的数量不足，从而降低分类算法的精度，如'love'，该单词满足统计特征，但由于其词性为动词，因此该单词无法被系统提取。

图 1(a)表明，当单词满足八种统计特征时，即在标准 C_8 时，基于多统计特征的情感词提取算法具有最高的精度，当单词至少满足一种统计特征时，即在标准 C_1 时，提取算法的精度最低。由表 9-(a)可知，基于 C_1 构造的情感词典在商品评论分类测试中具有最高的精度，而基于 C_8 构造的情感词典在商品评论分类测试中具有最低的精度。造成该现象的原因在于，由于提取标准 C_i 要求被提取的单词至少满足 i 种统计特征，当 i 增加时，满足要求的单词也随之减少，候选词情感词列表中只有少量单词满足标准 C_8，从而造成在分类测试中词典 D8 无法提供足够数量的情感词，降低分类器的准确率。例如，'cheap'在商品评论中该单词能够表达客户对商品价格的观点，在基于多统计特征的提取实验中，该单词满足三种特征，因此只有词典 D1~D3 包含'cheap'，当使用这些词典进行商品评论分类时，可能会遗漏包含'cheap'的文本，影响分类算法的效率。

对比表 8(a)与表 9(a)可知，在朴素贝叶斯分类器中基于单一统计特征的商品评论分类算法优于基于多统计特征的文本情感分类算法，而在另外四种分类器中，基于多统计特征的文本情感分类算法具有更高的精度。并且，在多统计特征的商品评论分类实验中，基于 8 种统计特征创建向量空间模型，与传统的基于单词构造向量空间模型的方法相比，该方法有效的降低了文本向量的维度，具有隐性语义空间(LSA/SDV)的压缩效果，压缩文本向量可以有效减小数据的规模，降低了分类算法的空

间和时间复杂度。

5　结束语

本文选取了 8 种在自然语言处理中常见的统计特征，并研究它们在情感词抽取和商品评论分类中的作用。实验结果表明，基于统计特征的文本情感分类方法具有更高的精度。在基于多统计特征的商品评论分类实验中，以 8 种统计特征为基础创建文本的向量空间模型，替代传统的文本表示方法。测试结果表明这文本表示方法在保证分类算法准确率和召回率的前提下，有效的降低了分类算法的时间和空间复杂度。

今后的工作将改进语句拆分算法，使系统可以挖掘文本中包含的网络用语，研究统计特征在不同的分类算法中的权重，使系统能够根据分类器自动为对应的统计特征赋予相应的权重，提高文本分类的效率。

参考文献：

[1] Tang D Y, Wei F R, Qin B, et al. Building large-scale twitter-specific sentiment lexicon: a representation learning approach [C]// Proc of the 25th International Conference on Computational Linguistics. 2014: 23-29.

[2] Ibrahim H S, Sherif M A, Gheith M H. Sentiment analysis for modern standard Arabic and colloquial [J]. International Journal on Natural Language Computing, 2015, 4 (2): 95-109.

[3] Wang F X, Zhang Z H, Lan M. ECNU at SemEval-2016 task 7: an enhanced supervised learning method for lexicon sentiment intensity ranking [C]// Proc of the International Workshop on Semantic Evaluation. 2016: 491-496.

[4] Mohammad S M, Bravo-Marquez F. Wassa-2017 shared task on emotion intensity [C]// Proc of the Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis. 2017.

[5] Qiu G, Liu B, Bu J J, et al. Opinion word expansion and target extraction through double propagation [J]. Computational Linguistics, 2011, 37 (1): 9-27.

[6] Liu K, Xu L H, Zhao J. Extracting opinion targets and opinion words from online reviews with graph co-ranking [C]// Proc of the 52nd Annual Meeting of the Association for Computational Linguistics. 2014: 314-324.

[7] Chetviorkin I, Loukachevitch N. Domex: extraction of sentiment lexicons for domains and meta-domains [C]// Proc of COLING. 2012: 77-86.

[8] Jovanoski D, Pachovski V, Nakov P. On the impact of seed words on sentiment polarity lexicon induction [C]// Proc of COLING. 2016: 11-17.

[9] Severyn A, Moschitti A. On the automatic learning of sentiment lexicons [C]// Proc of Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2015: 1397-1402.

[10] Yu H L, Deng Z H, Li S Y. Identifying sentiment words using an optimization-based model without seedwords [C]// Proc of the 51st Annual Meeting of the Association for Computational Linguistics. 2013: 855-859.

[11] Rajeswari K, Nakil S, Patil N, et al. Text categorization optimization by a hybrid approach using multiple feature selection and feature extraction methods [J]. International Journal of Engineering Research and Applications, 2014, 4 (3): 86-90.

[12] Uysal A K. An improved global feature selection scheme for text classification [J]. Expert Systems With Applications, 2016, 43: 82-92.

[13] McAuley J, Leskovec J. Hidden factors and hidden topics: understanding rating dimensions with review Text [C]// Proc of the 7th ACM Conference on Recommender Systems. 2013: 165-172.

[14] Chen Z Y, Arjun Mukherjee, Liu B. Aspect extraction with automated prior knowledge learning [C]// Proc of the 52nd Annual Meeting of the Association for Computational Linguistics. 2014, 347-358.

[15] Mesleh A A. Chi square feature extraction based svms arabic language text categorization system [J]. Journal of Computer Science, 2007, 3 (6): 430-435.

[16] Mitra P, Murthy C A, Sankar K. P. Unsupervised feature selection using feature similarity [J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2002, 24 (3): 301-312.

[17] Juola P, Baayen H. A controlled-corpus experiment in authorship identification by cross-entropy [J]. Literary and Linguistic Computing, 2005, 20 (1): 59-67.

[18] Wang Y, Witten I. Inducing model trees for continuous classes [C]// Proc of the 9th European Conference on Machine Learning. 1997: 128-137.

[19] Zhou X J, Wan X J, Xiao J G. Collective opinion target extraction in Chinese microblogs [C]// Proc of Conference on Empirical Methods in Natural Language Processing. 2013: 1840-1850.

[20] Bakliwal A, Arora P, Patil A, et al. Towards enhanced opinion classification using NLP techniques [C]// Proc of Workshop on Sentiment Analysis where AI Meets. 2011: 101-107.

[21] Yoo J Y, Yang D M. Classification scheme of unstructured text document using tf-idf and naïve bayes classifier [J]. Advanced Science and Technology Letters, 2015, 111 (50): 263-266.

[22] Chen H, Zhan Y, Li Y. The application of decision tree in Chinese email classification [C]// Proc of the 9th International Conference on machine Learning and Cybernetics. 2010: 305-308.

[23] Zhang M L, Zhou Z H. Multi-label neural networks with applications to functional genomics and text categorization [J]. IEEE Trans on Knowledge and Data Engineering, 2006, 18 (10): 1338-1351.

[24] Moreira S, Filgueiras J, Martins B, et al. Reaction: a naive machine learning approach for sentiment classification [C]// Proc of the 2nd Joint Conference on Lexical and Computational Semantics. 2013: 490-494.

[25] Pang B, Lee L, Shivakumar Vaithyanathan. Thumbs up?sentiment classification using machine learning techniques [C]// Proc of Conference on Empirical Methods in Natural Language Processing. 2002: 79-86.